CLASSIFICATION OF THYROID NODULES IN ULTRASOUND IMAGES USING DEEP MODEL BASED TRANSFER LEARNING AND HYBRID FEATURES

Tianjiao Liu[1], Shuaining Xie[1], Jing Yu[2], Lijuan Niu[3], Weidong Sun[1]

[1]Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, P. R. China
[2]Colg. of Computer Science & Technology, Beijing Univ. of Technology, Beijing 100124, P. R. China
[3]Cancer Hospital of Chinese Academy of Medical Sciences, Beijing 100021, P. R. China

ABSTRACT

Ultrasonography is a valuable diagnosis method for thyroid nodules. Automatically discriminating benign and malignant nodules in the ultrasound images can provide aided diagnosis suggestions, or increase the diagnosis accuracy when lack of experts. The core problem in this issue is how to capture appropriate features for this specific task. Here, we propose a feature extraction method for ultrasound images based on the convolution neural networks (CNNs), try to introduce more meaningful semantic features to the classification. Firstly, a CNN model trained with a massive natural dataset is transferred to the ultrasound image domain, to generate semantic deep features and handle the small sample problem. Then, we combine those deep features with conventional features such as Histogram of Oriented Gradient (HOG) and Local Binary Patterns (LBP) together, to form a hybrid feature space. Finally, a positive-sample-first majority voting and a feature-selected based strategy are employed for the hybrid classification. Experimental results on 1037 images show that the accuracy of our proposed method is 0.931, which outperformed other relative methods by over 10%.

Index Terms— classification, transfer learning, feature fusion, deep learning, ultrasound image

1. INTRODUCTION

Thyroid nodule is one of the most common endocrine carcinoma. Ultrasonography has become the most widely used modality for detecting and diagnosing thyroid cancer, for its better reveal ability and distinction between benign and malignant nodules in pathological features compared to CT and MRI. With the rapid development of medical imaging technology, computer aided diagnosis (CAD) assists us solve the subjective diagnosing problem existing in current method, which highly depends on personal experience. As a well-trained 'expert', it has a wide application prospect in any situation when need to close the experience gap. A fully-automatic CAD progress includes image preprocessing like denoising, ROI extraction and classification. Recently, much attention has been focused on the first two phases, while works on classification using basic ultrasound images are still rare, especially when restricted to thyroid nodule classification. The challenge of such classification problem mainly lies in how to select distinguishable features, thus most researches have focused on the feature design of various types, such as morphometric features and texture features. In [1], Zakeri et al. proposed some effective texture features for differentiating breast nodules. Ding et al. combined B-mode image and elastogram to gain both local texture features and global elasticity features [2], and Owjimehr et al. chose completed LBP texture to classify liver images [3]. But, experimental results reveal that these features perform quite unsatisfactorily on our dataset, and the reason probably arises from the internal simplicity and locality of low-level features. The built-in disadvantages of ultrasound image such as speckle noises and low contrast, along with the variations in shape, size, and trait of thyroid nodules, incapacitate those features for their working well in such classification task. In clinical diagnosis, doctors concerning more pathological features at semantic layer, also verifies such argument. Thus, high-level features with semantic meaning should be induced to obtain better classification quality.

Deep learning models, specifically CNNs, are widely used and performed extremely well in various visual recognition tasks such as object detection and image classification [4,5]. Features extracted by CNNs can be seen as complex hierarchical representations of the inputs, which can well capture recessive characteristics inside the images. These high-level semantic features are exactly what we need as supplement for low-level features. However, those CNNs consist of millions of nodes and configurations, which means only large datasets can support the training progress. This bottleneck lay obstacles for applications in

Acknowledgement: This work was supported by the Capital Health Research and Development of Special (No.2014-2-4025) of China.

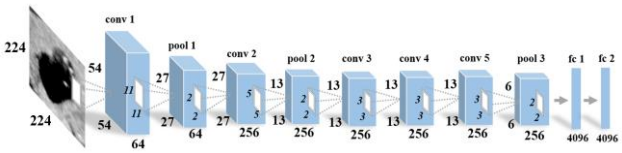

Fig.1. The architecture of the VGG model used in this work.

the medical field, for such large datasets are usually unavailable. Initial studies have demonstrated that transfer learning can overcome this problem by applying the well-trained deep models to other image datasets for feature extraction, since CNNs do not aim to handle a specific problem, but learn to capture the inherent characteristics of visual objects [6-11].

In this paper, we transfer the CNNs model learned from ImageNet [5] as a pre-trained feature extractor to our ultrasound image dataset, to explore its versatility as a high-level feature description. To get better classification results, we propose a hybrid approach combining traditional low-level features with deep learning features for thyroid nodule classification. To the best of our knowledge, this is the first attempt to transfer deep learning features to the ultrasound images for thyroid nodule classification, and the results of our proposed method demonstrate remarkable improvement in the classification accuracy.

2. THE PROPOSED METHOD

2.1. Pre-trained VGG-F Model for Transfer Learning

CNNs are a class of deep learning models, from which the high-level characteristics can be extracted. CNNs adopt a feed-forward working mode, where the features generated by the former layer are inputted in the intermediate layer, and the outputs are then passed to the next layer [12]. In the classification task, CNN can be used as a feature extractor, which may be transferred to some distinct but related problems. Although the appearance of ultrasound images is quite different from natural images, the way of acknowledging features is the same. The more samples are trained, the more universal the features are. ImageNet provides a large database for training deep models, thus in this work, we employ the VGG-F model trained from ImageNet to extract features and complete classification.

VGG-F model, designed by an Oxford group [13], is consist of 5 convolutional-pooling groups and 2 fully-connected FC layers as shown in Fig.1. Both fully-connected layers have 4096-dimentional outputs that can be directly used as feature descriptor for classification. During this processing procedure, the output values of each layer can be considered as features to some extent. A visualize result is shown in Fig. 2 to explain how an image response to a certain convolution layer, in other words, a group of filters. The results exported from the lower layers, share the analogous low-level features, including edge, directional

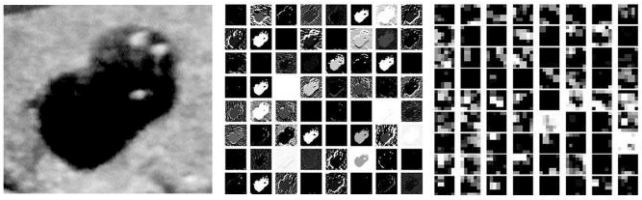

Fig. 2. Response of a certain convolution layer: (a) Input image. (b) Outputs of the 3rd layer. (c) Part of outputs of the 7th layer.

and intensity features, with those extracted by using Gabor filter. While, for the images outputted from the last several layers, various features compositely emerge. Thus, in order to identify the most discriminative features for the nodule classification tasks, we compare the features extracted from the pool3, fc1 and fc2 layers of the VGG network.

2.2. Feature fusion and Classification

Except the generalization high-level feature descriptors from the CNNs, handcrafted low-level feature descriptors still retain the advantages in pertinence. We have explored several traditional low-level features for handling our ultrasound images: (1) Gray level co-occurrence matrix (GLCM) [14]; (2) Local binary patterns (LBP), which applied for thyroid tissue texture in [15]; (3) Histogram of oriented gradient (HOG), which used for automatic view classification of ultrasound echocardiogram images [16]); (4) Scale-invariant feature transform (SIFT) and vector of locally aggregated descriptors (VLAD).

For each features describing image characteristics from different aspects, the combination of those features may lead to a more comprehensive representation of pathological characteristics of nodules. Among these features, we integrated HOG, SIFT and LBP features with the high-level features extracted from CNNs and jointed them to form a one-dimensional vector. The redundant elements in the feature vector should be removed so as to avoid overfitting. To this end, two kinds of feature fusion strategies are used in our proposed method:

1) Feature-selected strategy: Firstly, extracting the features, and directly connecting those features into one feature vector and reducing the dimension. The dimension of the features extracted from the VGG-F model is 4096. Adding with 144-dimensional HOG feature, 26-dimensional LBP feature, and 512-dimensional VLAD feature, the final dimension of the feature vector is 4778. In order to reducing the redundancies and irrelevancies among feature vector, feature selection is required. The feature selection standard is based on sorting the differences of benign samples and malignant samples [5]:

$$diff_k = \left| \frac{1}{N_{MB}} \sum_{i=1}^{N_{MB}} v_{ik} - \frac{1}{N_{MM}} \sum_{i=1}^{N_{MM}} v_{ik} \right| (k = 1, \cdots, N) \qquad (1)$$

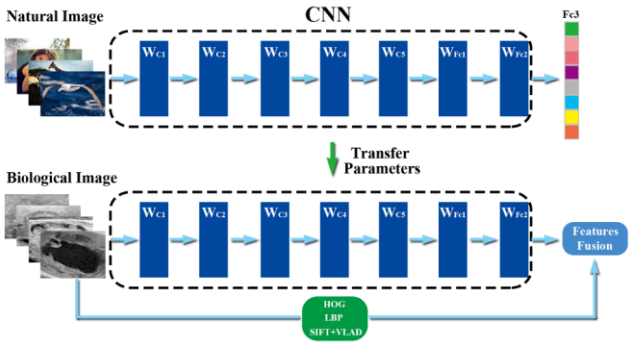

Fig. 3. Flowchart of deep models for transfer learning and feature fusion.

Where, N_{MB} and N_{MM} are the number of benign and malignant nodules in the training set, v_{ik} is the kth dimensional feature of the ith image. The top 1000 features with the largest differences will be chosen as our final features for the thyroid nodule classification.

2) Positive-sample-first majority voting strategy: Different features are separately imported in the classifier at the first stage, and then fuse all the predicted results given by the classifier under a designed decision strategy, and the final decision can be thus made. In this paper, we propose a simple and effective multiple feature fusion classification method based on the majority voting mechanism. Suppose T types of features can be extracted from the thyroid ultrasound images. For a feature extraction method k, a classifier h_k can be trained on the dataset. The value predicted by the classifier for sample $\boldsymbol{x}$ is $h_k(\boldsymbol{x})$. The final predicted classification result for sample $\boldsymbol{x}$ is expressed as follows, $h(x) = mode\ (h_1(x),...,h_T(x))$, where the *mode* is striving for the modal operation. If the votes of benign and malignance are the same on the condition that T is an even number, the sample is considered as malignancy. Once the same votes of benign and malignance occurs, we regard the corresponding results as malignance, based on the consideration of increasing the classification sensitivity. The algorithm flowchart of our method is demonstrated in Fig. 3.

3. EXPERIMENTAL RESULTS

3.1. Data and Evaluation Indexes

Thyroid ultrasound images used in our experiments are supplied by the Cancer Hospital of Chinese Academy of Medical Sciences, which are clinically verified. The ultrasound machine used in our experiments is GE Logiq E9, the frequency of detector is 10-14MHz. 1037 thyroid nodule ultrasound images, including 651 benign images and 386 malignant images, are used in our experiments and the type and location of all nodules are annotated by doctors.

We extracted the ROI of each nodule images by doctors' demarcation and the classification was performed using SVM classifier with 10-fold cross validation. Quantitative evaluation indexes are as follows: (1) *Accuracy*= $(TP+TN)/(TP+TN+FP+FN)$; (2) *Sensitivity*=$TP/(TP+FN)$; (3) *Specificity* =$TN/(TN+FP)$; (4) Receiver operating characteristic curve (ROC) and area under curve (AUC). Where *TP* (true positive) and *TN* (true negative) represent the positive and negative sample number of right classification, respectively. *FP* (false positive) and *FN* (false negative) are the negative and positive sample number of false classification. In the classification of the thyroid nodule samples, positive samples are the malignant nodules, and vice versa. The sensitivity and specificity define the possibility of predicting the malignant and benign nodules, respectively.

Tables 1. Comparison results of different features using 4 evaluation indexes.

Features	Accuracy	Sensitivity	Specificity	AUC
HOG	0.829	0.699	0.906	0.837
LBP	0.780	0.562	0.909	0.793
SIFT+VLAD	0.824	0.726	0.882	0.841
POOL3	0.856	0.773	0.905	0.917
FC1	0.853	0.774	0.911	0.940
FC2	0.860	0.788	0.911	0.946
POOL3+	**0.931**	**0.908**	**0.945**	**0.977**
FC1+	0.923	0.895	0.939	0.982
FC2+	0.920	0.887	0.940	0.976
Vote-POOL3+	0.918	0.902	0.928	0.963
Vote-FC1+	0.913	0.885	0.929	0.959
Votr-FC2+	0.917	0.888	0.934	0.956

3.2. Results and Discussion

Compared with HOG, LBP and SIFT, the deep features from CNN have the best performance in the overall experiments. A detailed comparison result of different features using 4 evaluation indexes is given in Tables 1, in which, 'POOL3', 'FC1' and 'FC2' denote the features extracted from layer pool3, fc1 and fc2; 'POOL3+' 'FC1+' and 'FC2+' denotes 1000-dimensional feature fusion method using feature selection; and 'Vote-' denotes feature fusion method using voting strategy. From this table, we find that specificity scores higher than sensitivity in every feature considered. It may result from two reasons: Firstly, the positive samples are nearly twice as many as the negative ones in our dataset. Secondly, the diversity of malignant tumors makes individual differences greater than benign nodules. Besides, benign nodules have more distinct characteristics, making it easier to be recognized.

A more intuitional result is shown in Fig.4(a). This indicates that highly hybrid features generated by CNNs can represent the inherent character of images regardless of the image type, thus it is feasible to transfer CNN features to ultrasound images domain. Besides, features generated from different CNN layers perform differently. Although they

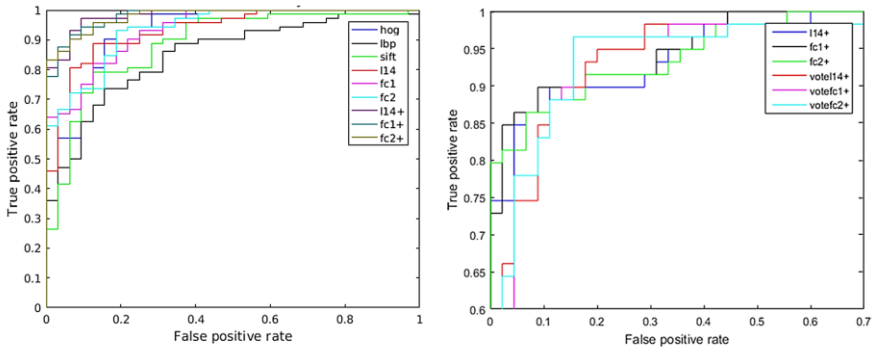

Fig.4. Comparative analysis results of ROC: (a) ROC on different features. (b) ROC between two fusion strategies.

perform similarly on accuracy, we can observe that AUC increases when layers go higher. Under the condition of the unbalanced data collection, AUC can persuasively embody the performance of classifiers. This relatively high AUC value indicates that fc2 is a superior classification feature group. One reasonable explanation is that higher layer capture features which are generic among different kind of images, while the lower ones only compute low-level features which cannot represent high-level semantics.

From Fig.4(a), we can also see that, our proposed feature fusing method achieves an accuracy of 93.1% when using L14 features. Compared with the deep features from CNN and the traditional low-level features, these two feature fusion methods can generally improve the accuracy by 7%-8% and 10%-14%, respectively. Based on the discussions above, we find that fusing different features can always contribute to satisfactory results. The fusion strategies perform fairly well on sensitivity that can be generally 11% increased, compared with non-fusion method. The feature-selected based strategy includes feature combining and screening, removing redundant and undesirable features. Based on these treatments, a group of features with maximal discrimination can be thus acquired. While, for voting mechanism, the sensitivity can be increased by focusing on the recognition of positive samples, and adopting positive-sample-inclined strategy. The differences between two strategies may be found in Fig. 4(b), for the feature-selected one aims to improve the TP rate when FP rate stays low while the other makes effort on ensuring the classification accuracy of positive samples.

4. CONCLUSION

In this paper, a feature extraction method for ultrasound images is presented to classify the thyroid nodules into benign and malignant. We considered both traditional low-level features and high-level deep features extracted from CNN model. Deep features can provide us more generic semantic meanings to the limited medical dataset. Among them, the fc2 layer generates the most representative ones. We have also compared the performance of our proposed hybrid method with the other related methods. The comparison results shown that, our proposed hybrid method outperformed both the pre-trained CNN model and the traditional single-type feature method. In future work, we plan to complete further tuning of the CNN, and improve classification accuracy especially for the malignant nodules.

Moreover, broader applications will be considered when we extend our method to other medical tasks.

5. REFERENCES

[1] F.S. Zakeri, H. Behnam, and N. Ahmadinejad, "Classification of Benign and Malignant Breast Masses Based on Shape and Texture Features in Sonography Images," *Journal of Medical Systems*, vol. 36, no.3, pp.1621-1627, 2012.

[2] J. Ding, et al., "Multiple-instance learning with global and local features for thyroid ultrasound image classification," in *BioMedical Engineering and Informatics* ,2014.

[3] M. Owjimehr, H. Danyali, and M.S. Helfroush, "Fully automatic segmentation and classification of liver ultrasound images using completed LBP texture features," in *International Conference on E-Business and E-Government*, 2014.

[4] Y. Gong, et al., "Multi-scale Orderless Pooling of Deep Convolutional Activation Features," in *European Conference on Computer Vision*, 2014.

[5] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Neural Information Processing Systems*, 2012.

[6] N. Jean, et al., "Combining satellite imagery and machine learning to predict poverty," *Science,* vol. 353, pp. 790-794, 2016.

[7] H. Ravishankar, et al., "Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning," in *International Symposium on Biomedical Imaging*, 2016.

[8] M. Oquab, et al., "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *Computer Vision and Pattern Recognition*, 2014.

[9] Y. Gao, M.A. Maraci, and J.A. Noble, "Describing ultrasound video content using deep convolutional neural networks," in *International Symposium on Biomedical Imaging*, 2016.
[10] Y. Xu, et al., "Deep convolutional activation features for large scale Brain Tumor histopathology image classification and segmentation," in *International Conference on Acoustics, Speech, and Signal Processing*, 2015.
[11] W. Zhang, et al., "Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis," in *Knowledge Discovery and Data Mining*, 2015.
[12] K. Chatfield, et al., "Return of the Devil in the Details: Delving Deep into Convolutional Nets," in *British Machine Vision Conference*, 2014.
[13] Y. Bar, et al., "Chest pathology detection using deep learning with non-medical training," in *International Symposium on Biomedical Imaging*, 2015.
[14] A.S.M. Sohail, et al., "Retrieval and classification of ultrasound images of ovarian cysts combining texture features and histogram moments," in *International Symposium on Biomedical Imaging*, 2010.
[15] E.G. Keramidas, et al., "Efficient and effective ultrasound image analysis scheme for thyroid nodule detection," in *International Conference on Image Analysis and Recognition*, 2007.
[16] D. Agarwal, K.S. Shriram, and N. Subramanian, "Automatic view classification of echocardiograms using histogram of oriented gradients," in *International Symposium on Biomedical Imaging,* 2013.